Classifications croisées de données de trajectoires contraintes

Mohamed Khalil El Mahrsi, Romain Guigourès, Fabrice Rossi, Marc Boullé

Classifications croisées de données de trajectoires contraintes par un réseau routier

Mohamed K. El Mahrsi[*,**], Romain Guigourès[**,***], Fabrice Rossi[**], Marc Boullé[***]

[*] Télécom ParisTech, Département Informatique et Réseaux
46, rue Barrault 75634 Paris CEDEX 13, France
khalil.mahrsi@telecom-paristech.fr
[**] Équipe SAMM EA 4543, Université Paris I Panthéon-Sorbonne
90, rue de Tolbiac 75634 Paris CEDEX 13, France
mohamed-khalil.el-mahrsi@univ-paris1.fr
romain.guigoures@univ-paris1.fr
fabrice.rossi@univ-paris1.fr
[***] Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion, France
romain.guigoures@orange.com
marc.boulle@orange.com

Résumé. Le clustering (ou classification non supervisée) de trajectoires a fait l'objet d'un nombre considérable de travaux de recherche. La majorité de ces travaux s'est intéressée au cas où les objets mobiles engendrant ces trajectoires se déplacent librement dans un espace euclidien et ne prennent pas en compte les contraintes liées à la structure sous-jacente du réseau qu'ils parcourent (ex. réseau routier). Dans le présent article, nous proposons au contraire la prise en compte explicite de ces contraintes. Nous représenterons les relations entre trajectoires et segments routiers par un graphe biparti et nous étudierons la classification de ses sommets. Nous illustrerons, sur un jeu de données synthétiques, l'utilité d'une telle étude pour comprendre la dynamique du mouvement dans le réseau routier et analyser le comportement des véhicules qui l'empruntent.

1 Introduction

Le monitoring du trafic routier est effectué, dans la majorité des cas, grâce à des capteurs dédiés qui permettent d'estimer le nombre de véhicules traversant la portion routière sur laquelle ils sont installés. Les coûts prohibitifs d'installation et de maintenance pour ce genre de capteurs limitent leur déploiement au réseau routier primaire (c.à-d. les autoroutes et les grandes artères seulement). Par conséquent, ce genre de solutions produit une information incomplète sur l'état du réseau routier, ce qui complique l'extraction de connaissances sur la dynamique des mouvements dans ce réseau et sur l'adéquation entre le réseau et son usage.

Une solution alternative (ou complémentaire) consiste à exploiter des traces GPS d'objets mobiles recueillies par des dispositifs ad hoc (par exemple des smartphones). Ces traces peuvent être obtenues lors de campagnes d'acquisition spécifiques (bus, taxis, flotte d'entreprise, etc.)

ou par des mécanismes de crowdsourcing en proposant à des utilisateurs de soumettre leurs propres trajets. On peut ainsi obtenir un volume important d'information couvrant le réseau de façon beaucoup plus complète que des capteurs.

Le clustering (ou classification non supervisée) figure parmi les techniques d'analyse les plus utiles à de telles fins exploratoires. La majorité des travaux traitant du clustering de trajectoires s'est focalisée sur le cas du mouvement libre (Nanni et Pedreschi, 2006), (Benkert et al., 2006), (Lee et al., 2007), (Jeung et al., 2008) en faisant abstraction des contraintes liées à la topologie du réseau routier, qui jouent pourtant un grand rôle dans la caractérisation de la similarité entre les trajectoires analysées. Parmi les travaux ayant traité le cas contraint (Kharrat et al., 2008) ; (Roh et Hwang, 2010). El Mahrsi et Rossi (2012b) proposent de représenter les relations entre différentes trajectoires sous forme d'un graphe et de s'intéresser au clustering de ce dernier pour découvrir des groupes de trajectoires de profils similaires. Les auteurs étendent ce travail dans (El Mahrsi et Rossi, 2012a) en s'intéressant aux regroupements de segments routiers – toujours en se basant sur une représentation par graphes – afin d'enrichir la connaissances des groupes de trajectoires et d'apporter un moyen supplémentaire de les interpréter. Nous proposons, dans cet article, de conserver cette représentation des données sous la forme d'un graphe. Plus précisément, nous modéliserons les relations qu'entretiennent les trajectoires et les segments routiers sous forme d'un graphe biparti et nous étudierons deux approches différentes de classification de ses sommets.

Le reste de l'article est organisé comme suit. La section 2 présente notre modèle de données ainsi que les approches que nous proposons. Section 3 illustre notre étude expérimentale et démontre l'intérêt de ces approches et leur capacité à mettre en valeur des structures de clusters intéressantes tant au niveau des trajectoires qu'au niveau des segments routiers. Enfin, une conclusion sera dressée dans la section 4.

2 Approches de classification

Dans le cas contraint, une trajectoire T est modélisée sous forme d'une succession de segments routiers appartenant à l'ensemble de tous les segments constituant le réseau. Nous modélisons les données sous forme d'un graphe biparti $\mathcal{G} = (\mathcal{T}, \mathcal{S}, \mathcal{E})$. $\mathcal{T}$ est l'ensemble des trajectoires, $\mathcal{S}$ est l'ensemble de tous les segments du réseau routier et $\mathcal{E}$ est l'ensemble des arêtes modélisant les passages des trajectoires de $\mathcal{T}$ sur les segments de $\mathcal{S}$.

Dans un premier temps, nous proposons de projeter le graphe $\mathcal{G}$ afin d'étudier séparément les graphes correspondant respectivement aux trajectoires d'une part et aux segments d'autre part (Section 2.1). Dans un second temps, le graphe biparti est traité directement grâce à une approche de biclustering (Section 2.2).

2.1 Approche par projections de graphes

La projection du graphe $\mathcal{G}$ sur l'ensemble de ses sommets représentant les trajectoires $\mathcal{T}$ produit un graphe $\mathcal{G}_{\mathcal{T}} = (\mathcal{T}, \mathcal{E}_{\mathcal{T}}, \mathcal{W}_{\mathcal{T}})$, décrivant les relations de similarité entre les trajectoires. Une arête $e_{\langle T_i, T_j \rangle}$ relie deux trajectoires T_i et T_j si celles-ci partagent au moins un segment routier.

La pondération $\omega_{\langle T_i, T_j \rangle}$ la plus basique de cette arête peut consister en un comptage des segments communs entre les deux trajectoires. Si nous pondérons avec la mesure de similarité

proposée par El Mahrsi et Rossi (2012b) la projection coïncide avec la définition du graphe de similarité entre trajectoires, introduite par les mêmes auteurs. C'est cette stratégie de pondération que nous adoptons par la suite. Le poids $\omega_{\langle T_i,T_j\rangle}$ est donc la similarité cosinus entre les deux trajectoires T_i et T_j exprimée comme suit :

$$\omega_{\langle T_i,T_j\rangle} = \frac{\sum_{s\in\mathcal{S}} w_{s,T_i} \cdot w_{s,T_j}}{\sqrt{\sum_{s\in\mathcal{S}} w_{s,T_i}^2} \cdot \sqrt{\sum_{s\in\mathcal{S}} w_{s,T_j}^2}}$$

Où $w_{s,T} = \frac{n_{s,T}\cdot\text{length}(s)}{\sum_{s'\in T} n_{s',T}\cdot\text{length}(s')} \cdot \log \frac{|\mathcal{T}|}{|\{T_i : s\in T_i\}|}$ est un tf-idf modifié attribué à chaque segment routier s en fonction de sa longueur, son importance dans la trajectoire T et sa fréquence dans le jeu de données.

De façon analogue, la projection du graphe $\mathcal{G}$ sur l'ensemble des segments $\mathcal{S}$ produit le graphe $\mathcal{G}_\mathcal{S} = (\mathcal{S}, \mathcal{E}_\mathcal{S}, \mathcal{W}_\mathcal{S})$ décrivant les relations entre segments routiers. Ici, une arête $e_{\langle s_i,s_j\rangle}$ relie deux segments s'il y a au moins une trajectoire qui les visite tous les deux. Là également, nous opterons pour la pondération proposée par El Mahrsi et Rossi (2012a) pour affecter les poids $\mathcal{W}_\mathcal{S}$ au lieu d'un comptage simple des trajectoires communes.

Nous nous proposons d'effectuer le clustering de chacun de ces deux graphes de façon isolée (c.à-d. chacun est traité à part) pour obtenir une classification des trajectoires et une des segments. Pour ce faire, nous utiliserons l'algorithme de détection de communautés dans les graphes par optimisation de la modularité préconisé par Noack et Rotta (2009). Ce choix – qui est motivé par la tendance de ces graphes à avoir des sommets à fort degré et par l'efficacité des approches basées sur la modularité dans ce cas précis – n'écarte pas la possibilité d'utiliser d'autres algorithmes de clustering de graphes tels que le clustering spectral (Meila et Shi, 2000) ou le clustering par propagation de labels (Raghavan et al., 2007).

Pour un jeu de données composé de n trajectoires qui parcourent un réseau routier composé de m segments, la complexité algorithmique théorique pour effectuer le clustering de trajectoires est de $O(n^3)$ tandis que celle du clustering de segments est de $O(m^3)$ (Noack et Rotta, 2009). Cependant, les complexités observées en pratiques sont plutôt quadratiques.

Nous croiserons ensuite les deux classifications et essayerons d'interpréter chacune d'entre elles en fonction de l'autre.

2.2 Approche par biclustering

Nous proposons ici d'étudier directement le graphe sous sa forme bipartie $\mathcal{G} = (\mathcal{T}, \mathcal{S}, \mathcal{E})$. Pour cela, nous appliquons une approche de biclustering sur la matrice d'adjacence du graphe : les segments sont représentés en colonnes, les trajectoires en lignes et l'intersection d'une ligne et d'une colonne indique le nombre de passages d'une trajectoire sur un segment. Le but d'un biclustering est de réordonner les lignes et les colonnes de manière à faire apparaitre et à extraire des blocs de densités homogènes dans la matrice d'adjacence du graphe biparti $\mathcal{G}$. Une fois ces blocs extraits, on en déduit deux partitions obtenues simultanément, une de segments et une de trajectoires.

Une structure de biclustering, que nous notons $\mathcal{M}$, est définie par un ensemble de paramètres de modélisations décrits dans le Tableau 1. Le but d'un algorithme de biclustering va être d'inférer la meilleure partition du graphe.

Graphe $\mathcal{G}$	**Modèle de biclustering $\mathcal{M}$**
$\mathcal{T}$: ensemble des trajectoires	$C_\mathcal{T}$: ensemble des clusters de trajectoires
$\mathcal{S}$: ensemble des segments	$C_\mathcal{S}$: ensemble des clusters de segments
$\mathcal{E} = \mathcal{T} \cap \mathcal{S}$: ensemble des passages des trajectoires sur les segments	$C_\mathcal{E} = C_\mathcal{T} \cap C_\mathcal{S}$: biclusters de trajectoires et de segments

TAB. 1 – *Notations.*

En appliquant ce type d'approches, les trajectoires sont regroupées si elles parcourent des segments communs et les segments sont regroupés s'ils sont parcourus par des trajectoires communes. L'avantage de cette technique est qu'elle ne requière pas de pré-traitement sur les données, ni de définition de mesure de similarité entre trajectoires ou entre segments. L'inconvénient principal réside dans la complexité algorithmique de ce type d'approches qui peut s'avérer très élevée.

Nous choisirons ici d'utiliser l'approche MODL (Boullé, 2011) afin d'inférer notre structure de biclustering. Cette approche non-paramétrique a des capacités de passage à l'échelle nous permettant de l'utiliser pour le problème que nous traitons dans cet article. Un critère est construit suivant une approche MAP (Maximum A Posteriori) :

$$\mathcal{M}^* = \underset{\mathcal{M}}{\operatorname{argmax}}\, P(\mathcal{M})P(\mathcal{D}|\mathcal{M}).$$

D'abord, une probabilité a priori $P(\mathcal{M})$ dépendant des données est définie. Elle spécifie les paramètres de modélisation en attribuant à chacun d'eux une pénalisation correspondant à leur longueur de codage minimale, obtenue grâce aux statistiques descriptives des données. Ainsi, plus une structure de biclustering sera parcimonieuse, moins elle sera coûteuse. Ensuite, la vraisemblance des données connaissant le modèle $P(\mathcal{D}|\mathcal{M})$ est définie. Elle mesure le coût de recodage des données $\mathcal{D}$ avec les paramètres du modèle $\mathcal{M}$. Donc, le modèle de biclustering le plus probable est le modèle le plus fidèle aux données initiales. En d'autres termes, la vraisemblance favorise les structures informatives. La définition du critère global est donc un compromis entre une structure de biclustering simple et synthétique, et une structure fine et informative.

D'un point de vue algorithmique, l'optimisation est réalisée à l'aide d'une heuristique gloutonne ascendante, initialisée avec le modèle le plus fin, c'est-à-dire avec un segment et une trajectoire par cluster. Elle considère toutes les fusions entre les clusters et réalise la meilleure d'entre elles si cette dernière permet de faire décroitre le critère optimisé. Cette heuristique est améliorée avec une étape de post-optimisation, pendant laquelle on effectue des permutations au sein des clusters. Le tout est englobé dans une métaheuristique de type VNS (Variable Neighborhood Search, Hansen et Mladenovic (2001)) qui tire profit de plusieurs lancements de l'algorithme avec des initialisations aléatoires différentes. L'algorithme est détaillé et évalué dans Boullé (2011).

La complexité algorithmique est en $\mathcal{O}(|\mathcal{E}|\sqrt{|\mathcal{E}|}\log(|\mathcal{E}|))$ avec $|\mathcal{E}|$ le nombre d'arcs du graphe biparti $\mathcal{G}$, qui correspondent, dans le cas présent, au nombre de passages de trajectoires sur les segments. Cette complexité est calculée au pire des cas, c'est-à-dire lorsque chaque trajectoire couvre chaque segment (matrice d'adjacence du graphe biparti pleine). En pratique,

l'algorithme est capable d'exploiter l'aspect creux habituellement observé dans ce type de données.

3 Étude expérimentale

Nous décrivons les données utilisées dans cette étude dans la section 3.1. Les résultats obtenus et leur interprétation sont donnés dans la section 3.2 et la section 3.3.

3.1 Données utilisées

Afin de tester notre proposition, nous utilisons des jeux de données synthétiques étiquetées (c.à-d. générées de façon à contenir des clusters de trajectoires qui sont supposés être les clusters naturels par la suite). La stratégie de génération de ces données est la suivante. L'espace couvert par le réseau routier (le rectangle minimal englobant tous ses sommets) est quadrillé en grille contenant des zones rectangulaires de tailles égales. Un cluster de trajectoires est alors généré comme suit. Une zone dans la grille du réseau routier est sélectionnée au hasard. Tous les sommets inclus dans cette zone sont sélectionnés pour jouer le rôle de points de départ éventuels pour les trajectoires appartenant au cluster. De façon similaire, une deuxième zone est sélectionnée au hasard et ses sommets sont retenus pour jouer le rôle de points d'arrivée. Pour chaque trajectoire à inclure dans le cluster, un sommet de départ (resp. d'arrivée) est tiré au hasard parmi les sommets de départ (resp. d'arrivée). La trajectoire est générée comme étant le plus court chemin reliant les deux sommets sélectionnés. Le nombre de trajectoires dans chaque cluster est fixé au hasard entre deux seuils paramétrables.

Pour illustrer les différentes informations qu'on peut tirer avec l'approche proposée et pour des soucis de clarté et de visibilité nous nous contentons de montrer les résultats obtenus sur un jeu de données composé de 85 trajectoires seulement. Ces trajectoires sont répandues sur cinq clusters distincts (cf. FIG. 1) et ont visité un total de 485 segments routiers distincts. Le jeu de données a été généré en utilisant la carte d'Oldenburg dont le graphe est composées de 6105 sommets et environ 14070 arcs.

3.2 Analyse des clusters de trajectoires

Le clustering par optimisation de la modularité du graphe des trajectoires produit, au départ, un partitionnement contenant trois clusters seulement et ne détecte donc pas les clusters naturels présents dans les données. Ce problème de résolution est d'ailleurs l'une des limitations des approches basées sur la modularité où certaines communautés restent fusionnées et ne sont donc pas détectées. Cependant, l'implémentation que nous utilisons (celle décrite dans Rossi et Villa-Vialaneix (2011)) résout ce phénomène en effectuant une descente récursive sur les communautés découvertes et produit donc une hiérarchie de clusters emboités. Le deuxième niveau de cette hiérarchie révèle l'existence de huit clusters. La matrice croisée de ceux-ci avec les clusters naturels est illustrée dans TAB. 2 qui montre que les clusters trouvés sont purs. Trois des clusters originaux ont été retrouvés de façon exacte tandis que les deux autres ont été éclatées sur plusieurs clusters plus fins (le cluster 1 est éclaté en trois classes et le cluster 3 sur deux). Ce choix de "sur-partitionnement" reste, cependant, tout à fait légitime et justifiable au vu des différences assez notables entre trajectoires constituant chacun de ces deux clusters.

(a) Cluster 1 (14 trajectoires) (b) Cluster 2 (19 trajectoires) (c) Cluster 3 (20 trajectoires)

(d) Cluster 4 (20 trajectoires) (e) Cluster 5 (12 trajectoires)

FIG. 1 – *Clusters "naturels" de trajectoires dans le jeu de données.*

Le biclustering génère une partition des trajectoires fidèle aux motifs générés aléatoirement. La matrice de confusion (Tableau 3) montre que les classes de trajectoires retrouvées par le biclustering sont pures, seules deux classes artificielles ont été scindées en deux par la méthode MODL. Cette technique sera donc préférée puisqu'elle découvre des motifs similaires aux motifs obtenus par maximisation de modularité, en étudiant directement le graphe biparti, se passant ainsi de toute projection et pré-traitements.

3.3 Analyse croisée des clusters

Nous proposons maintenant d'étudier la matrice d'adjacence du graphe biparti d'origine. On a réordonné les lignes et les colonnes de cette matrice de manière à rapprocher les trajectoires et les segments regroupés dans les mêmes clusters (voir Figure 2).

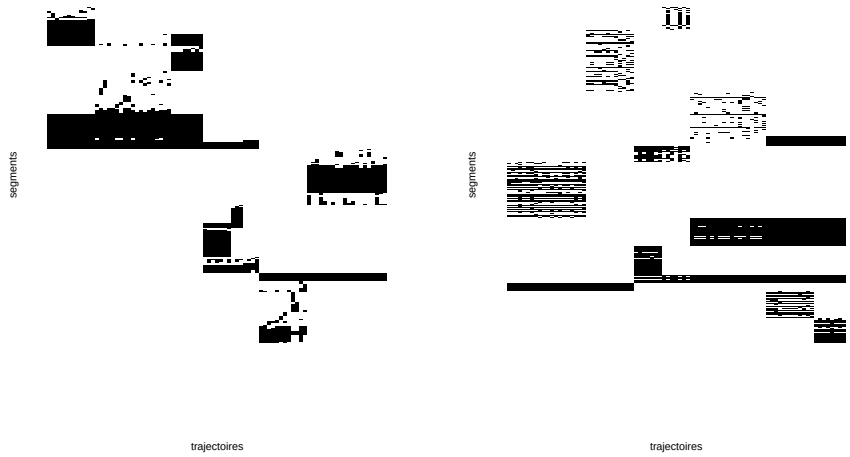
On observe dans le cas de l'étude de graphes projetés (Figure 2(a)) que les clusters regroupent des segments parcourus par les mêmes trajectoires, peu importe la quantité de trafic supportée. Les segments peu empruntés seront donc rattachés aux segments très empruntés par les rares trajectoires communes. Cela se caractérise dans la matrice par la présence de cellules (intersection des clusters de trajectoires et de segments) avec des distributions hétérogènes : certains segments sont couverts par toutes les trajectoires, d'autres ne sont parcourus que par quelques trajectoires.

	1	2	3	4	5	6	7	8
1	0	0	0	7	3	4	0	0
2	0	19	0	0	0	0	0	0
3	12	0	8	0	0	0	0	0
4	0	0	0	0	0	0	0	20
5	0	0	0	0	0	0	12	0

TAB. 2 – *Matrice de confusion entre clusters naturels (sur les lignes) et ceux obtenus par optimisation de la modularité (sur les colonnes).*

	1	2	3	4	5	6	7
1	0	0	7	7	0	0	0
2	0	0	0	0	19	0	0
3	0	0	0	0	0	12	8
4	20	0	0	0	0	0	0
5	0	12	0	0	0	0	0

TAB. 3 – *Matrice de confusion entre clusters naturels de trajectoires (sur les lignes) et clusters optimaux obtenus par biclustering (sur les colonnes).*



(a) Modularité (b) Biclustering

FIG. 2 – *Matrices d'adjacence des clusters croisés.*

A contrario, les clusters de segments obtenus par biclustering sont corrélés avec leur usage. On va donc pouvoir caractériser ces usages dans le réseau et ainsi détecter les *hubs* (Figure 3(a)),

les axes secondaires (Figure 3(b)) ou encore les ruelles peu empruntées. Le résultat obtenu ici est donc une caractérisation de la structure topologique sous-jacente du réseau, dont l'information sur les usages est apportée par les trajectoires. Cela se matérialise sur la Figure 2(b) par des cellules de densité homogènes.

(a) Hub autoroutier

(b) Axes secondaires menant vers la périphérie de la ville

FIG. 3 – *Exemple de clusters de segments.*

L'information mutuelle est une mesure fréquemment utilisée en biclustering. Elle permet de quantifier les corrélations entre les partitions des deux variables étudiées, ici les segments et les trajectoires. L'information mutuelle est toujours positive et est d'autant plus importante que les clusters de trajectoires parcourent des clusters de segments uniques. Ici nous proposons d'étudier la contribution à l'information mutuelle. Il s'agit de quantifier l'apport d'un couple de clusters de trajectoires et de clusters de segments sur l'information mutuelle du modèle.

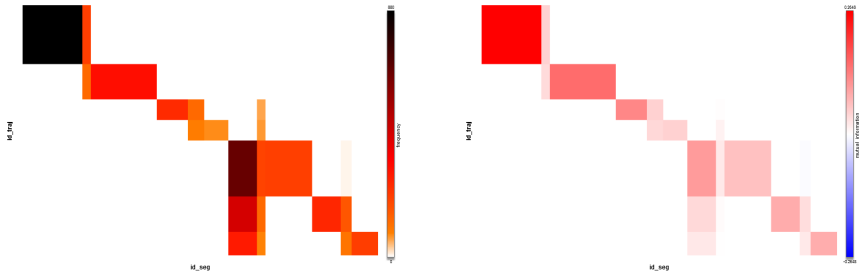
Définition (Contribution à l'information mutuelle)**.** *La contribution à l'information mutuelle, notée* $mi(c_S, c_T)$*, est définie de la manière suivante :*

$$mi(c_S, c_T) = P(c_S, c_T) \log \frac{P(c_S, c_T)}{P(c_S)P(c_T)} \tag{1}$$

où $P(c_S, c_T)$ *est le probabilité pour un passage d'appartenir à une trajectoire de* c_T *et de couvrir un segment de* c_S*,* $P(c_S)$ *est la probabilité de parcourir un segment du cluster* c_S *et* $P(c_T)$*, la probabilité d'être sur une trajectoire de* c_T*.*

Une contribution positive à l'information mutuelle signifie que le nombre de passages des trajectoires du cluster c_T sur les segments du cluster c_S est supérieur à la quantité de trafic attendu en cas d'indépendance des clusters de trajectoire et de de segments. Dans le cas d'une contribution négative, on observe une quantité de trafic inférieure à la quantité attendue. Enfin, une contribution à l'information mutuelle nulle montre une quantité attendue de trafic ou alors un trafic très faible ou nul.

La Figure 4(b) présente les contributions à l'information mutuelle de chaque couple de biclusters. Le bicluster en haut à gauche est très caractéristique dans le sens où le cluster de segments n'est traversé que par un cluster de trajectoires et le cluster de trajectoire passe



(a) Fréquences des biclusters de segments et trajectoires

(b) Information mutuelle entre clusters de trajectoires et de segments

FIG. 4 – *Fréquence et information mutuelle pour les biclusters découverts.*

principalement par ce cluster de segments. Dans le cas présent, le cluster de trajectoires contient $21{,}6\%$ des trajectoires étudiées et le cluster de segments $17{,}3\%$ des segments du jeu de données. On s'attend donc, en cas d'indépendance, à observer $21{,}6\% \times 17{,}3\% = 3{,}7\%$ des parcours totaux. Or ici, on observe $17{,}3\%$ du parcours totaux, ce qui représente un important excès de trafic sur le groupe de segments par le groupe de trajectoires, par rapport au résultat attendu en cas d'indépendance.

L'information mutuelle présente une information différente de celle apportée par la matrice de fréquence. On observe sur certains clusters de segments, un nombre de parcours significatifs par plusieurs clusters de trajectoires. Ce type de clusters est caractéristique des *hubs* routiers. Certains de ces clusters présentent peu de contrastes en terme d'information mutuelle, ce qui signifie que, malgré la nature de hub du cluster, le trafic y est plutôt bien réparti.

4 Conclusion

Dans cet article nous avons étudié la classification des données de trajectoires sous un angle de clustering de graphes bipartis. L'apport principal de cette étude se situe sur le plan méthodologique où nous avons montré l'intérêt de ce genre d'approches pour extraire des connaissances utiles sur le comportement des usagers du réseau routier. Nous avons notamment étudié le problème, dans un premier ordre, comme étant un problème de détection de communautés dans deux graphes séparés décrivant les trajectoires d'une part et les segments routiers d'une autre part. Nous avons, ensuite, étudié le biclustering direct du graphe biparti décrivant les trajectoires et les segments en même temps. Les algorithmes de clustering utilisés ici (par optimisation de la modularité dans les cas des projections du graphe biparti et MODL pour le biclustering) servent à illustrer l'intérêt de notre formulation du problème. Il est donc tout à fait possible de les remplacer par d'autres algorithmes de clustering de graphes tels que le clustering spectral.

Il serait intéressant de tester nos approches sur des données réelles et d'en comparer les résultats avec des faits réels. Il est également intéressant d'étudier leur comportement en présence de données bruitées où les clusters à découvrir sont moins évidents.

Références

Benkert, M., J. Gudmundsson, F. Hübner, et T. Wolle (2006). Reporting flock patterns. In *ESA'06 : Proceedings of the 14th conference on Annual European Symposium*, London, UK, pp. 660–671. Springer-Verlag.

Boullé, M. (2011). Data grid models for preparation and modeling in supervised learning. In *Hands-On Pattern Recognition : Challenges in Machine Learning, vol. 1*, pp. 99–130. Microtome.

El Mahrsi, M. K. et F. Rossi (2012a). Graph-Based Approaches to Clustering Network-Constrained Trajectory Data. In *Proceedings of the Workshop on New Frontiers in Mining Complex Patterns (NFMCP 2012)*, Bristol, Royaume-Uni, pp. 184–195.

El Mahrsi, M. K. et F. Rossi (2012b). Modularity-Based Clustering for Network-Constrained Trajectories. In *Proceedings of the 20-th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2012)*, Bruges, Belgique, pp. 471–476.

Hansen, P. et N. Mladenovic (2001). Variable neighborhood search : Principles and applications. *European Journal of Operational Research 130*(3), 449–467.

Jeung, H., H. T. Shen, et X. Zhou (2008). Convoy queries in spatio-temporal databases. In *ICDE '08 : Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, Washington, DC, USA, pp. 1457–1459. IEEE Computer Society.

Kharrat, A., I. S. Popa, K. Zeitouni, et S. Faiz (2008). Clustering algorithm for network constraint trajectories. In *SDH*, Lecture Notes in Geoinformation and Cartography, pp. 631–647. Springer.

Lee, J.-G., J. Han, et K.-Y. Whang (2007). Trajectory clustering : a partition-and-group framework. In *SIGMOD '07 : Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, New York, NY, USA, pp. 593–604. ACM.

Meila, M. et J. Shi (2000). Learning Segmentation by Random Walks. In *NIPS*, pp. 873–879.

Nanni, M. et D. Pedreschi (2006). Time-focused clustering of trajectories of moving objects. *J. Intell. Inf. Syst. 27*(3), 267–289.

Noack, A. et R. Rotta (2009). Multi-level algorithms for modularity clustering. In *Proceedings of the 8th International Symposium on Experimental Algorithms*, SEA '09, Berlin, Heidelberg, pp. 257–268. Springer-Verlag.

Raghavan, U. N., R. Albert, et S. Kumara (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E 76*(3).

Roh, G.-P. et S.-w. Hwang (2010). Nncluster : An efficient clustering algorithm for road network trajectories. In *Database Systems for Advanced Applications*, Volume 5982 of *Lecture Notes in Computer Science*, pp. 47–61. Springer Berlin - Heidelberg.

Rossi, F. et N. Villa-Vialaneix (2011). Représentation hiérarchique d'un grand réseau à partir d'une classification hiérarchique de ses sommets. *Journal de la Société Française de Statistique 152*, 34–65.

Summary

Trajectory clustering was studied mainly and extensively in the case where moving objects can move freely on the euclidean space. In this paper, we study the problem of clustering trajectories of vehicles whose movement is restricted by the underlying road network. We model relations between these trajectories and road segments as a bipartite graph and we try to cluster its vertices. We demonstrate our approaches on a synthetic dataset and show how it could be useful in inferring knowledge about the flow dynamics and the behavior of the drivers using the road network.